FELIX application in sPHENIX TPC

Jin Huang (BNL)

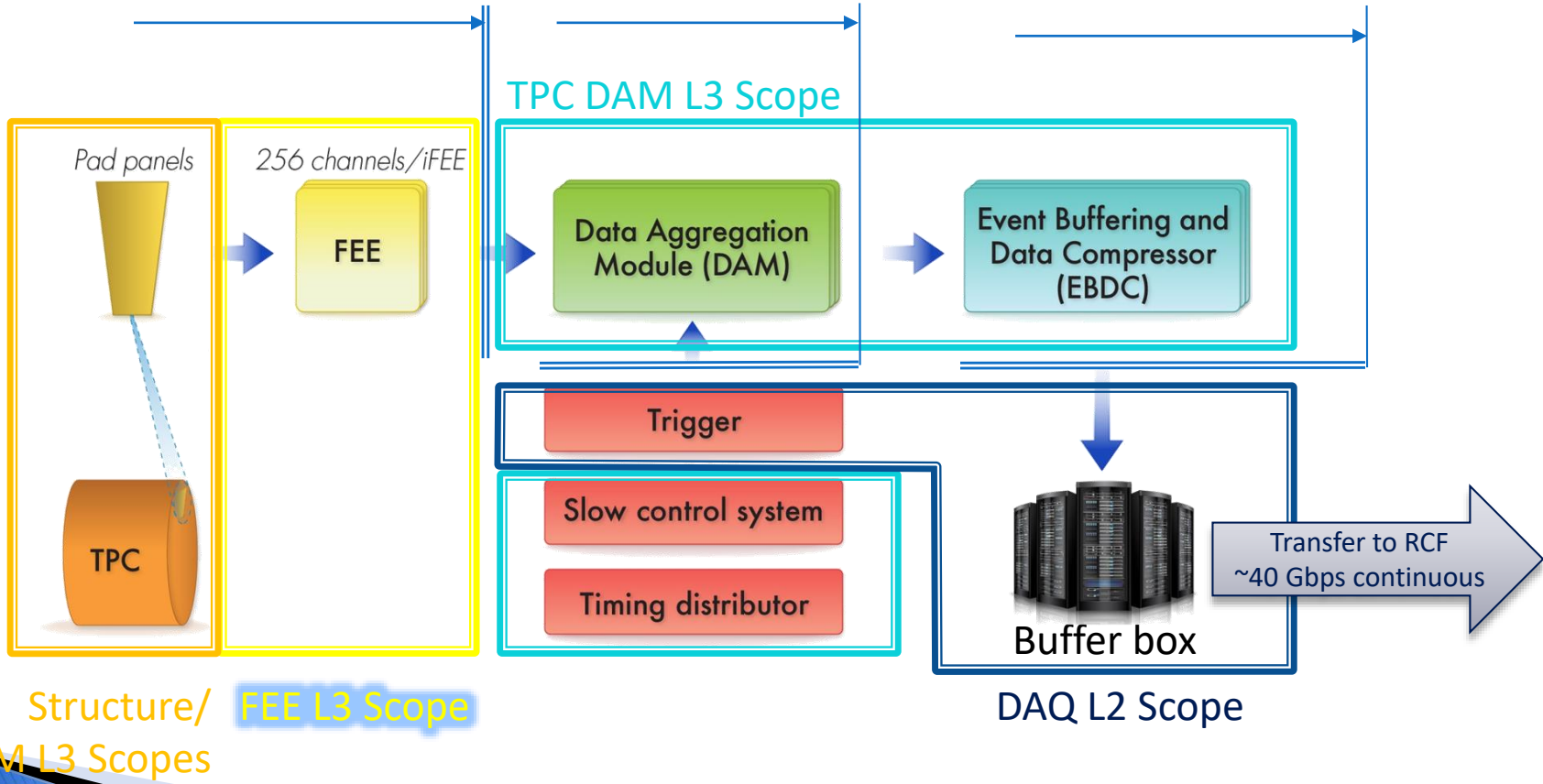
Envelop parameters for TPC DAM

Proposed KPP: demonstrate readout simulated data @ 1.56 Gbps x 600 fibers @ >99% LT

Input data stream:
600 4-Gbps fibers total
Max continuous: 2.87 Gbps / fiber
Average continuous: 1.43 Gbps x 600 fibers

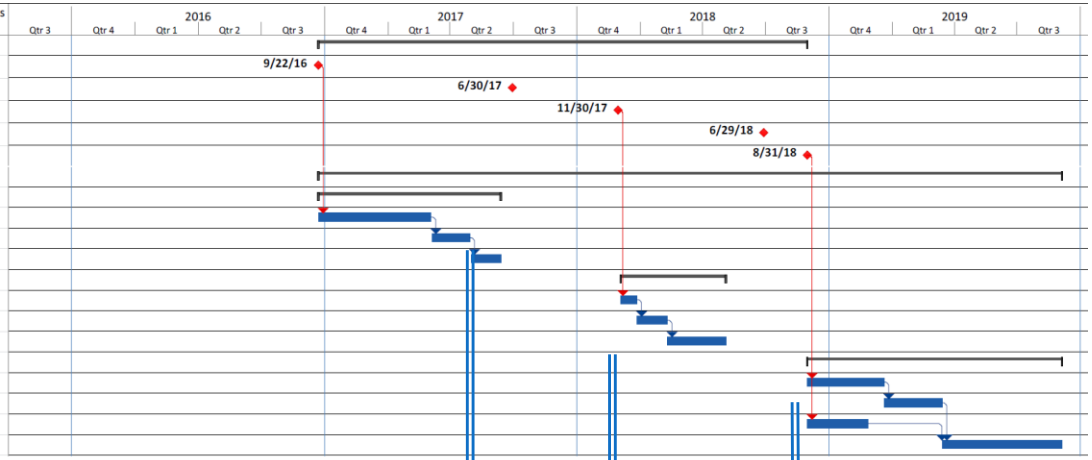
Clock/Trigger input:
Fiber, protocol TBD
Clock = 9.4 MHz
Trigger Rate = 15 kHz

Output data stream to buffer box:
N x 10 Gbps Ethernet via fiber (N=10-50)
Total continuous limit: <120 Gbps (?)
i.e. 3x (Transfer rate to RCF ~ 40 Gbps)



Timeline envelop. Cost ~ 0.5 M$

ID	WBS	Task Name	Duration	Start	Finish	Predecessors
1	**1.2.1**	**Tracker Management**	**484 days**	**Thu 9/22/16**	**Fri 8/31/18**	
2	1.2.1.1	CD-0	0 days	Thu 9/22/16	Thu 9/22/16	
3	1.2.1.2	CD-1, CD-2a/CD-3a Review	0 days	Fri 6/30/17	Fri 6/30/17	
4	1.2.1.3	CD-1, CD-2a/CD-3a Authorization	0 days	Thu 11/30/17	Thu 11/30/17	
5	1.2.1.4	CD-2b/CD-3b Review	0 days	Fri 6/29/18	Fri 6/29/18	
6	1.2.1.5	CD-2b/CD-3b Authorization	0 days	Fri 8/31/18	Fri 8/31/18	
192	**1.2.7**	**TPC DAM**	**734 days**	**Thu 9/22/16**	**Tue 9/3/19**	
193	**1.2.7.1**	**TPC DAM Evaluation --FELIX 1.5**	**180 days**	**Thu 9/22/16**	**Mon 6/12/17**	
194	1.2.7.1.1	Write Design Spec prototype v1	110 days	Thu 9/22/16	Fri 3/3/17	2
195	1.2.7.1.2	DAM Test Stand Development	40 days	Mon 3/6/17	Fri 4/28/17	194
196	1.2.7.1.3	Throughput Measurement	30 days	Mon 5/1/17	Mon 6/12/17	195
197	**1.2.7.2**	**TPC DAM Evaluation -- FELIX 2.0**	**105 days**	**Mon 12/4/17**	**Fri 5/4/18**	
198	1.2.7.2.1	Procure FELIX 2.0 board	15 days	Mon 12/4/17	Tue 12/26/17	4
199	1.2.7.2.2	Throughput measurement	30 days	Wed 12/27/17	Thu 2/8/18	198
200	1.2.7.2.3	FPGA Algorithm development	60 days	Fri 2/9/18	Fri 5/4/18	199
201	**1.2.7.3**	**TPC DAM Production**	**250 days**	**Fri 8/31/18**	**Tue 9/3/19**	
202	1.2.7.3.1	Procure Felix 2.0 boards	75 days	Fri 8/31/18	Wed 12/19/18	6
203	1.2.7.3.2	QA TPC FELIX	55 days	Thu 12/20/18	Wed 3/13/19	202
204	1.2.7.3.3	Procurement of EDBC Computers and Peripherals	60 days	Fri 8/31/18	Mon 11/26/18	6
205	1.2.7.3.4	Test DAM System	120 days	Thu 3/14/19	Tue 9/3/19	204,203



Next Milestone

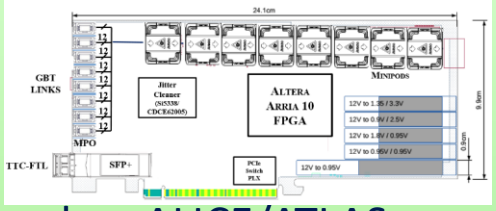
- Q4 2016, Design starts
- **June 2017, Feasible design, BNL CD1 review**
- **Nov 2017, DOE CD1 review**
- End 2017, Purchasing 2x FELIX v2.0 pre-production cards
- Mid 2018, CD3-b authorization, Procure 30 FELIX Cards
- Early 2020, Deliver all parts to 1008, establishing KPP
- Jan 2022, First beam, establishing UPP

Reminder: Full DAM system concept

Data Aggregation Module (DAM):
PCIex16 card with multiple (≥ 25) 3-Gbps fiber IO

Option 1: ATLAS FELIX

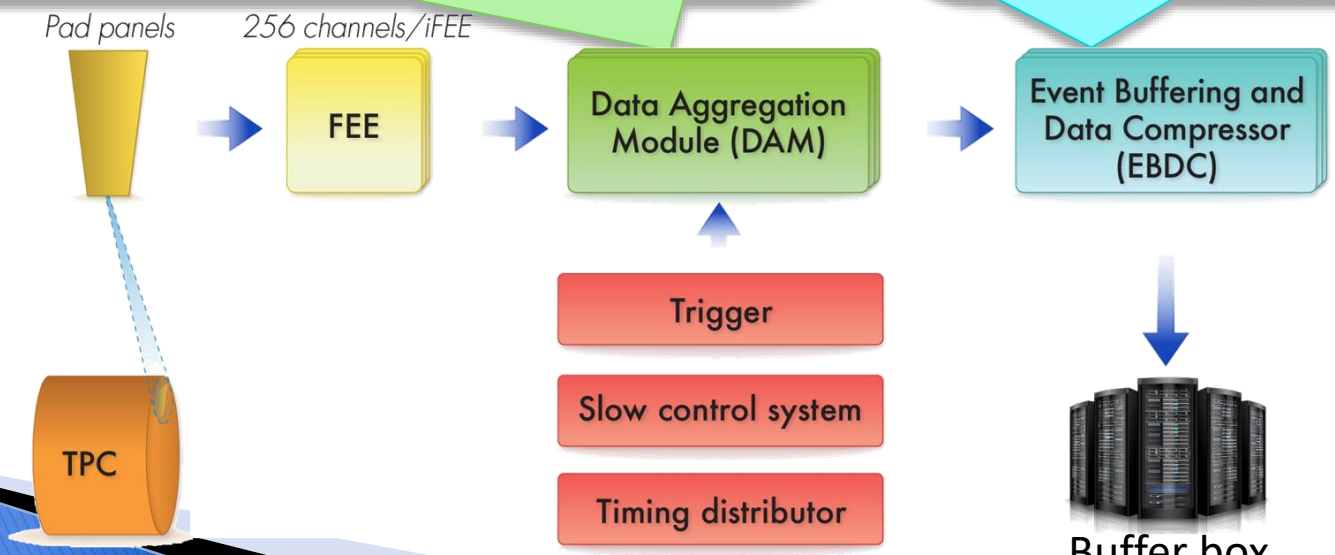
Option 2:LHCb/ALICE CRU



~~Option 3: build our own based on ALICE/ATLAS exp.~~

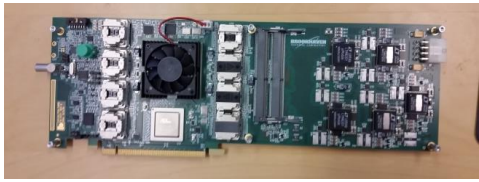
Event Buffering and Data Compressor (EBDC): Rack server that can host at 1x PCIex16 cards + 2x 10 Gbps Ethernet port

Example: Dell PowerEdge R830
12 cores, 1x10 GBps, ~ 6k$



FPGA Choices

(a) PCie40.



CRU/ PCIe40

FELIX

FPGA Family Name	Xilinx Virtex 6	Altera Stratix V GX	Xilinx Virtex 7	Altera Arria 10 GX **	Xilinx Virtex Ultrascale	Altera Stratix 10	CRU Requirements #	Xlinux Kintex Ultrascale
Status		available	available	ES available from Q2'15	available	end of 2017		**Available**
FPGA part number	XC6VLX240T	5SGXEA7	XC7VX690T	**10AX115**	XCVU190	10SG280		**XCKU115**
Used in	C-RORC	AMC40	MP7	**PCIe40**				**BNL 711, FELIX v1.5 prototype**
Logic Elements / Cells [M]	0.241	0.622	0.693	1.15	1.9	2.8		**1.451**
FFs [M]	0.3	0.939	0.866	1.7	2.14			**1.3**
LUTs [M]	0.15	0.235	0.433	0.425	1.07			**0.66**
18/20 Kb RAM Blocks	832	2560	2940	**2713**	7560	11721	**1920 / 2560**	**4320**
Total Block RAM (Mb)	15	50	53	**53**	133	229	**40 / 53**	**75.9**
≥ 10 Gb/s Transeivers	24	48	80	**96**	60	144	**48**	**(48 input + 48 output fiber links in FELIX)**
PLLs	12	28	20	32	60	48		**48**
PCIe x8, Gen3	2 (Gen2)	4	3	4	6	6		**6**

TPC Detector is the majority user (>70%) of CRU boards. CRU requirements is measured against TPC detector specific logic occupancy.

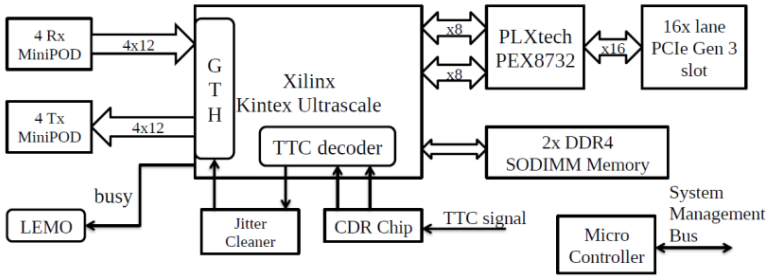
** Altough the maximum number of links of the Arria10 family is 96 links, the FPGA equiping the PCIe40 board has only 72 links

ATLAS group estimated 48x GBP unpacking and PCIe output use ~20% FELIX Resource

ATLAS/FELIX BNL-711 PCIe Card

Credit: Kai Chen (BNL), https://indico.bnl.gov/conferenceDisplay.py?confId=2653

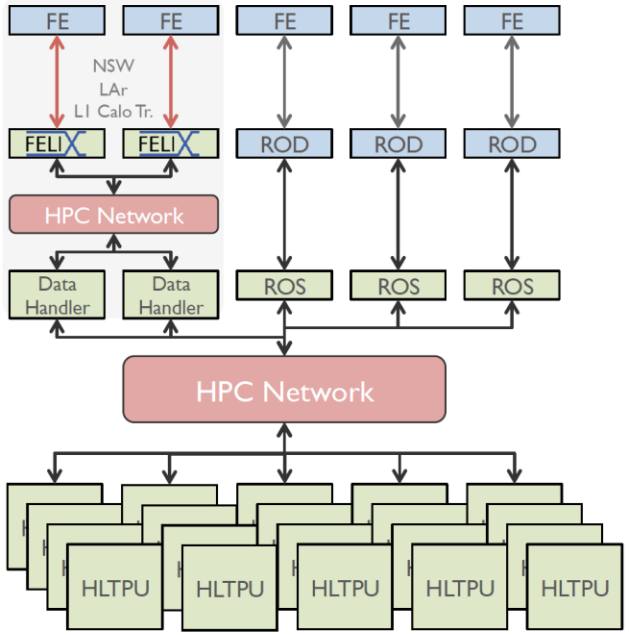
- BNL-711 Board chosen for ATLAS FELIX project, and used in ATLAS phase I upgrade, which is projected to complete before sPHENIX.
- Readout for ATLAS Phase-I sub-system of Liquid Argon Calorimeter, Level-1 calorimeter trigger, New small wheel of the muon spectrometer



Versatile Link, GBT (GigaBit Transceiver)

PCs

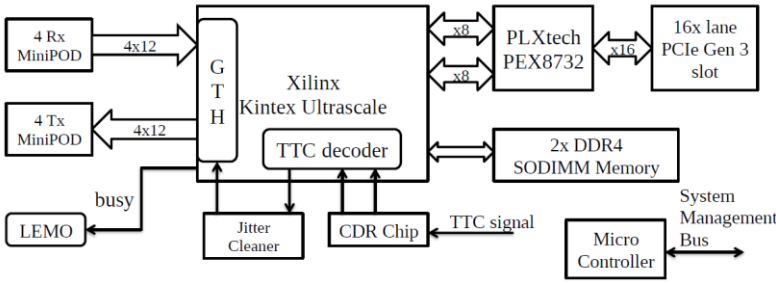
40 Gb Ethernet, Infiniband



ATLAS/FELIX Card for sPHENIX?

Credit: Kai Chen (BNL), https://indico.bnl.gov/conferenceDisplay.py?confId=2653

- Main features for FELIX PCIe Card
  - Design: BNL/Omega group, Layout: BNL/Instrumentation, Goal: multiple users.
  - A large Kintex Ultrascale FPGA, 1.5 M Logical Cells (24x Logical Cells of FVTX FEM card)
  - 48 bi-directional GBT link via two 48-F MTP connectors
  - PCIex16 Gen3, 101 Gbps demonstrated
  - 2x DDR4 memory slots (v1.0, v1.5), removed v2.0
  - TTC-timing input (v1.0, v1.5), timing mezzanine card (v2.0)

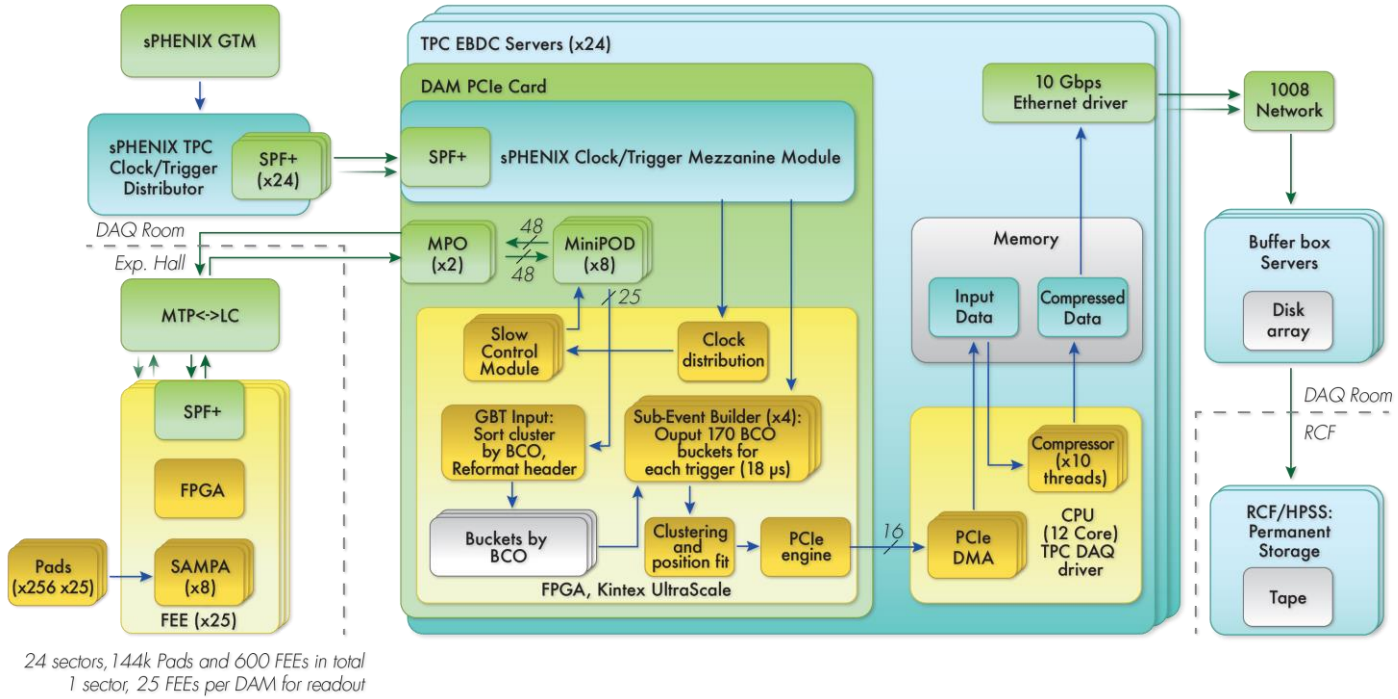


- Timeline and availability:
  - Current version: v1.5 prototype, can be ordered now
  - Next version: v2.0 pre-production, design completed, layout on-going, expect available Oct 2017
  - FELIX production system delivery expected end 2018 for ATLAS Phase-I upgrade. ATLAS needing 100+ card with various flavor of firmware depending on subsystem configurations.
- BNL/Omega group, Local expert expressed willing for help us to adapt FELIX in sPHENIX
  - Boards for initial evaluation test
  - Support firmware software development, timing mezzanine card design
  - The team is also help in possible use of FELIX card in proto-Dune.
  - The FELIX team is open for inputs in guiding the design to be more generic to various users.

FELIX v1.5 Card in test stand

Diagram & Rate

Live write up drafts online:
Rate estimation: https://docs.google.com/spreadsheets/d/1Q_uYf00_8pushSiYns29T_-ThIOqQaqpKbVS_LDqlAg/edit?usp=sharing
Data format: https://www.overleaf.com/read/wttbwnnynqwb



24 sectors, 144k Pads and 600 FEEs in total
1 sector, 25 FEEs per DAM for readout

Item			Rate Per unit		
	Unit	Count	Limit/Unit	Average/Unit	Max C./Unit
FEE SAMPA data	Gbps	4,800	1.28	0.20	0.51
FEE GBT fiber	Gbps	600	3.20	1.56	2.03
FPGA Input	Gbps	24	80.00	39.12	
Build hit - time table	Gbps	24	200.00	40.23	
After triggering	Gbps	24	200.00	11.47	
After clustering fitting	Gbps	24	101.70	5.73	
FPGA -> PCIex16 -> DMA	Gbps	24	101.70	5.73	
Lossless Compression	Gbps	24	4.80	3.44	
Server output to 1008 network	Gbps	24	10.00	3.44	
Buffer box servers	Gbps	1	120.00	82.56	

- 24 FELIX provide bi-direction link to 600 FEEs, 144k pads
  - Using 48x 48-F cable directly connect FELIX to TPC, breakup to LC in TPC
  - Input data to FELIX: 10MHz zero-suppressed wavelets @ 940 Gbps total
- Custom FPGA firmware
  - Buffer wavelets and sort into clock buckets
  - When triggered, group 18us of data or 170 buckets into events, reduce output data rate by factor of ~4.
  - Clustering and fitting: expect reduce data rate by factor of ~2
- On server CPU based compression then send to buffer box servers @ 80 Gbps

Recent works: FELIX v1.5 test stand

- Gain experience with operating a FELIX PCIe board
- **Demonstrate 4 Gbps bi-directional optical link, FEE <-> FELIX PCIe board (8b10b encoding)**
- Demonstrate top level FPGA design and resource counting
- Demonstrate FPGA -> PCIe -> DMA rate (Done. Already demonstrated by ATLAS group)

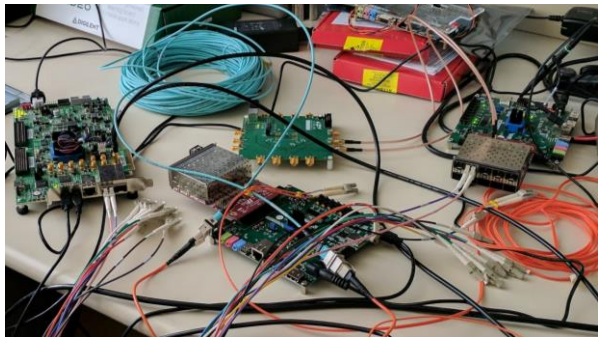
Xilinx Atrix-7 XC7A200T + 8x SFP

FELIX v1.5 PCIe card

Server

MTP -> LC

PCIex16, SMBus



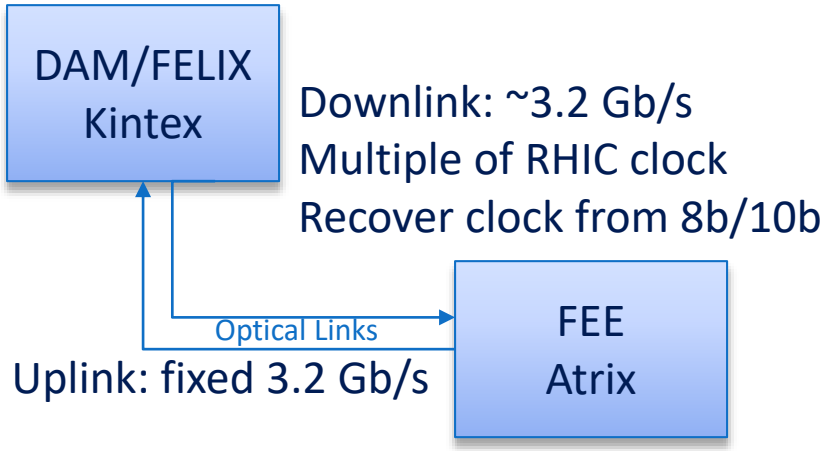
Also in test stand:

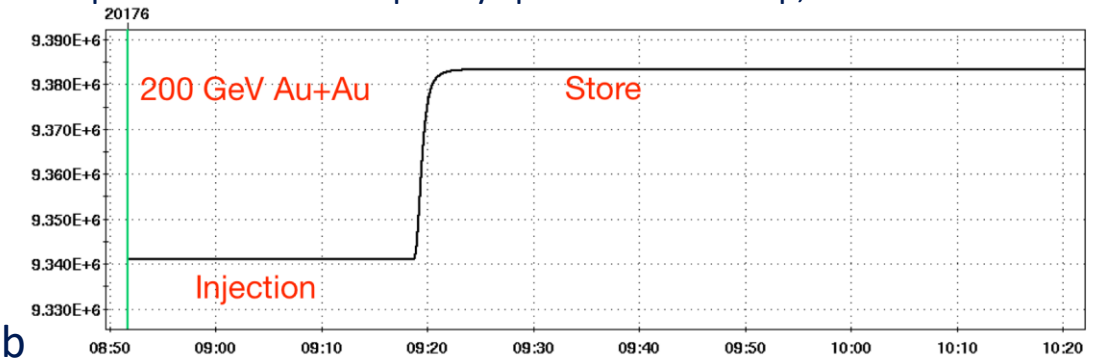
- Kintex evaluator board
- Si5338 PLL
- 50m – MTP fibers, etc.

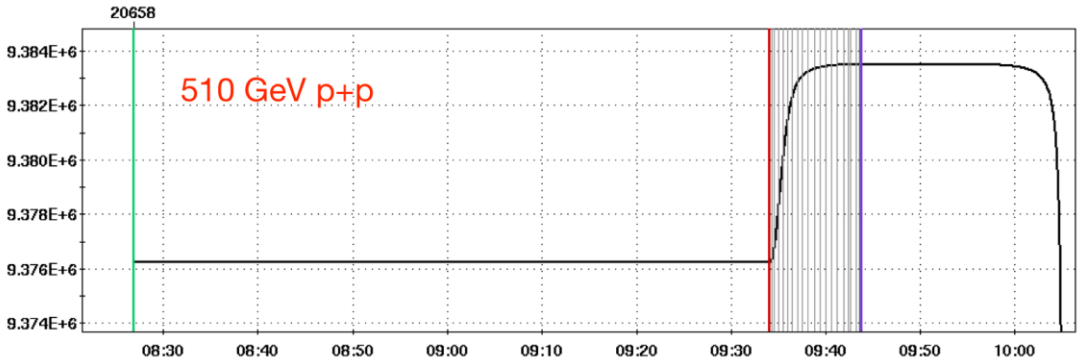
Kintex -> Atrix SFP+ fiber links with RHIC clock delivery

- One of the feature of the design is to use FELIX distribution slow control and RHIC clock to 25x FEE.
- Use RHIC clock (~9.4 MHz) as base frequency. Multiply x14 as fiber clock + slow control data in via 8b/10b encoder
- Recover RHIC clock on FEE, and slow control data via 8b/10b decoder
- John Kuczewski (BNL, instrumentation): tested with expected RHIC clock frequency spread (due to ramp). Excellent error rate

Special requirement: RHIC frequency spread due to ramp, 9.362 MHz ±22 kHz







Kintex -> Artix SFP+ fiber links

with RHIC clock delivery John Kuczewski (BNL, instrumentation)

Function generator mimic repeated RHIC clock ramping (triangle pattern)

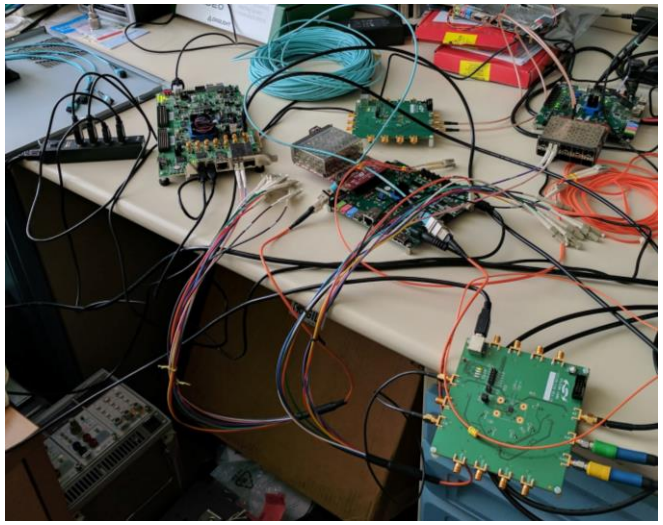
DAM/FELIX Kintex

Downlink: ~3.2 Gb/s
Multiple of RHIC clock
Recover clock from 8b/10b

Optical Links

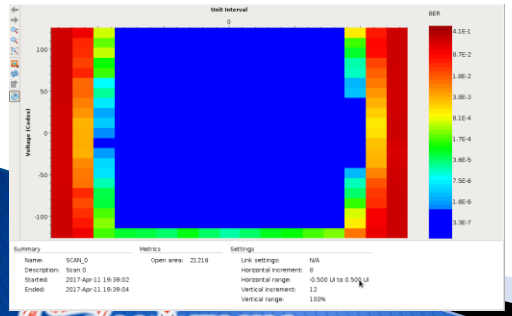
Uplink: 3.2 Gb/s from fixed clock

FEE Atrix

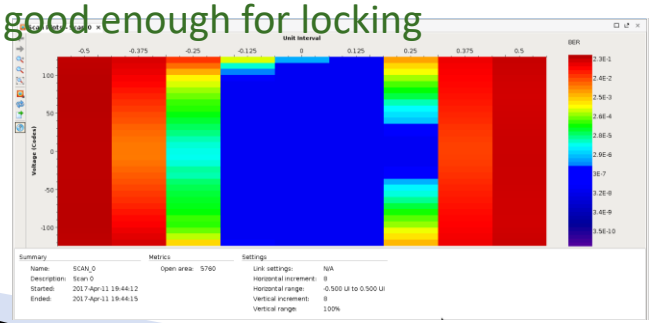


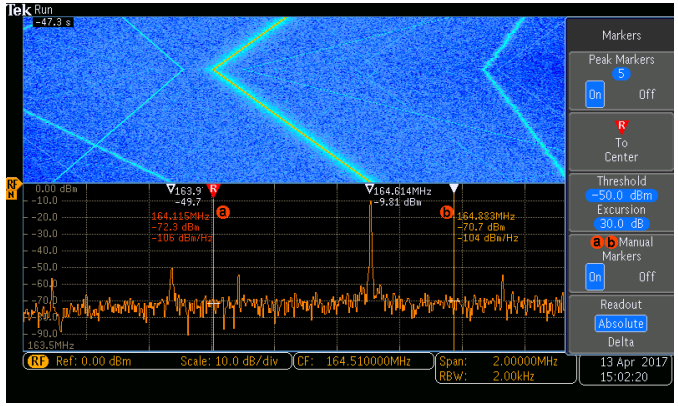
Recovered “RHIC” clock: tracking input pattern

Uplink “eye”: very stable



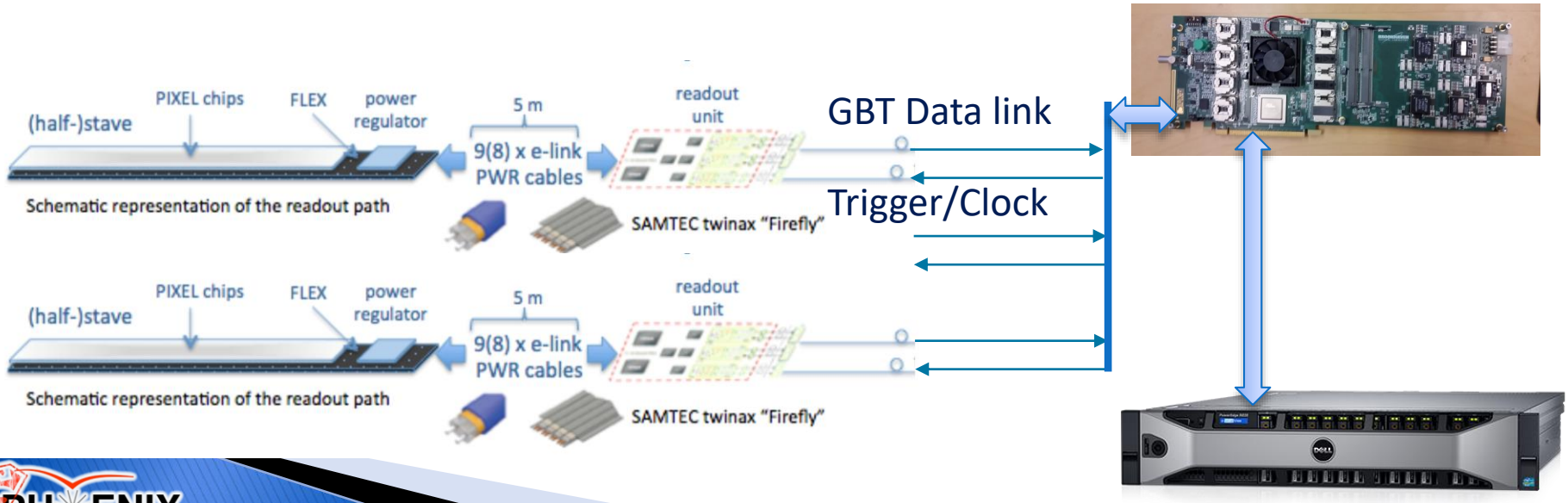
Downlink “eye” during ramping test: good enough for locking





TPC Outlook and possible use in MAPS

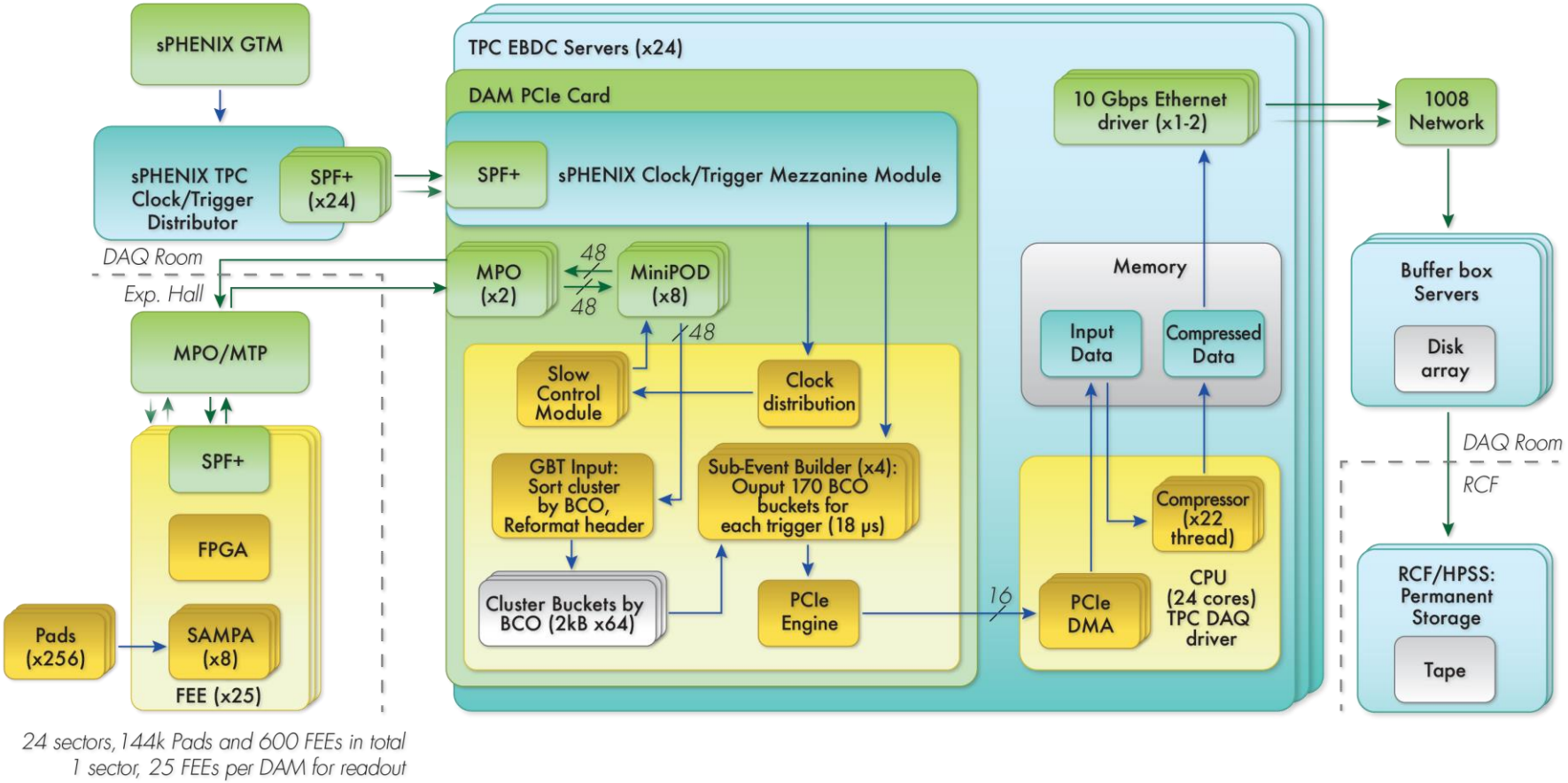
- The TPC group acquired v1.5 FELIX PCIe card to setup a test stand and evaluate DAQ feasibility. Plan to switch to pre-production cards end 2017. Meanwhile, we would also want to learn to CRU, prior to final decision on the readout cards.
- After Ming's request, a v1.5 FELIX PCIe card has been manufactured for MVTX.
- It makes sense to try pursuing the same system for MAPS+TPC readout, and share production batch, DAQ expertise and effort in timing distribution, card/server pool, event-building software development



Extra Information

DAM Plausibility diagram

Assuming 24x (DAM + EBDC) one for each TPC sector

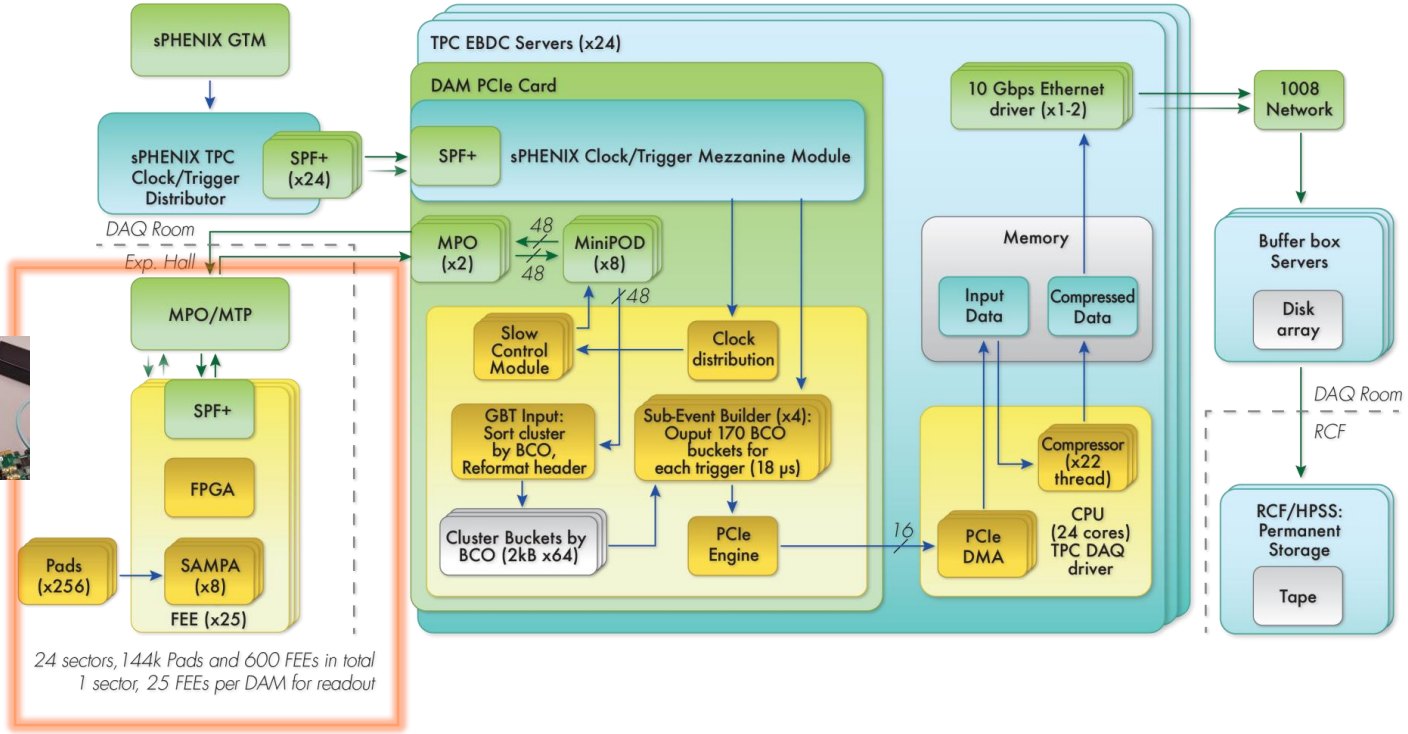


Rate estimation spread sheets:

https://docs.google.com/spreadsheets/d/1Q_uYf00_8pushSiYns29T_-ThIOqQaqpKbVS_LDqlAg/edit?usp=sharing

Input stage

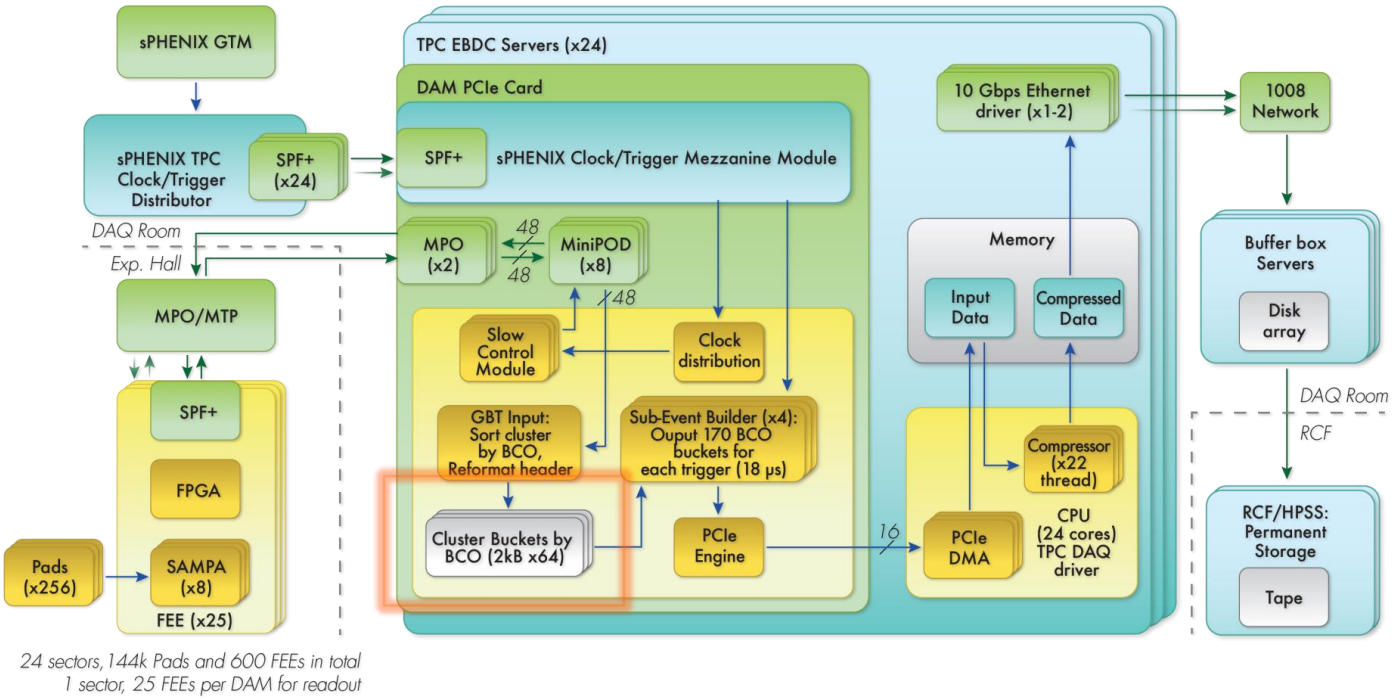
Rate estimation spread sheets:
https://docs.google.com/spreadsheets/d/1Q_uYf00_8pushSiYns29T_-ThIOqQaqpKbVS_LDqIAg/edit?usp=sharing



- Per DAM: ~25 FEEs, each send data in 1 fiber
  - Data format in minimal chunk = one cluster in one channel: 2x10 bit header (channel ID + timing + length) + 5x10 bit wavelet
  - Wavelet sampled timed to BCO (beam collision clock = 9.4 MHz)
  - **Max continuous rate / fiber = 1.9 Gbps, Average continuous rate / DAM = ~1 Gbps x 25 = 24 Gbps (30 pad rows)**
- Media: MTP fiber bundle, split to LC connector near detector. 8b/10b protocol?
- Downlink fiber send clock and slow control to FEEs. One DAM->FEE downlink fiber per FEE.
- Each FEE uniquely address by DAM ID (fixed by ID config in EBDC server) + DAM/FEE channel ID (fixed by cable mapping of DAM -> FEE)

BCO buckets

Rate estimation spread sheets:
https://docs.google.com/spreadsheets/d/1Q_uYf00_8pushSiYns29T_-ThIOqQaqpKbVS_LDqIAg/edit?usp=sharing

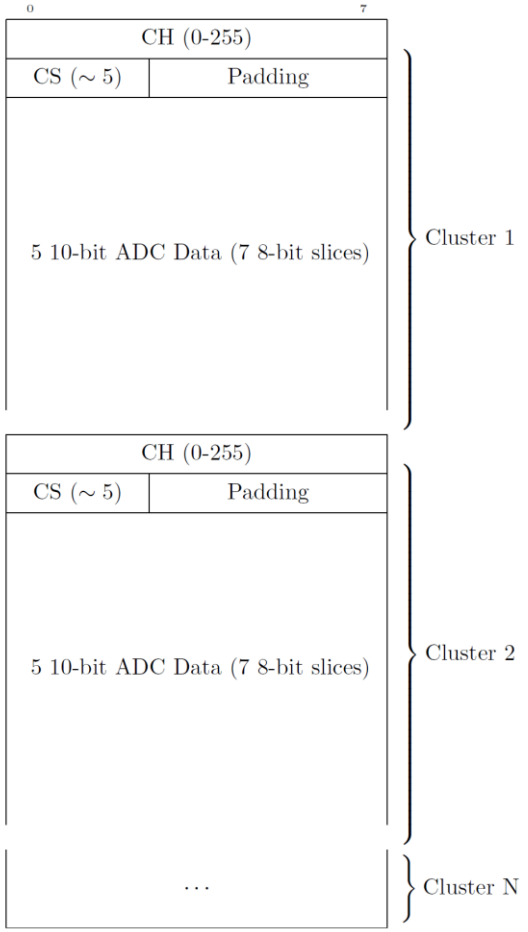


- FPGA are separated into 25 copies of DAM/FEE process channels, each handle one FEE input
- Separate clusters into buckets
  - Data format in minimal chunk = one cluster in one channel:
    2x10 bit header (channel ID + length) + 5x10 bit wavelet
  - Buffer long enough to allow transmission time spread, FVTX used 64 BCO buckets
  - Use internal memory on FPGA for BCO buckets storage
- Average continuous rate = 25 Gbps **(30 pad rows)**

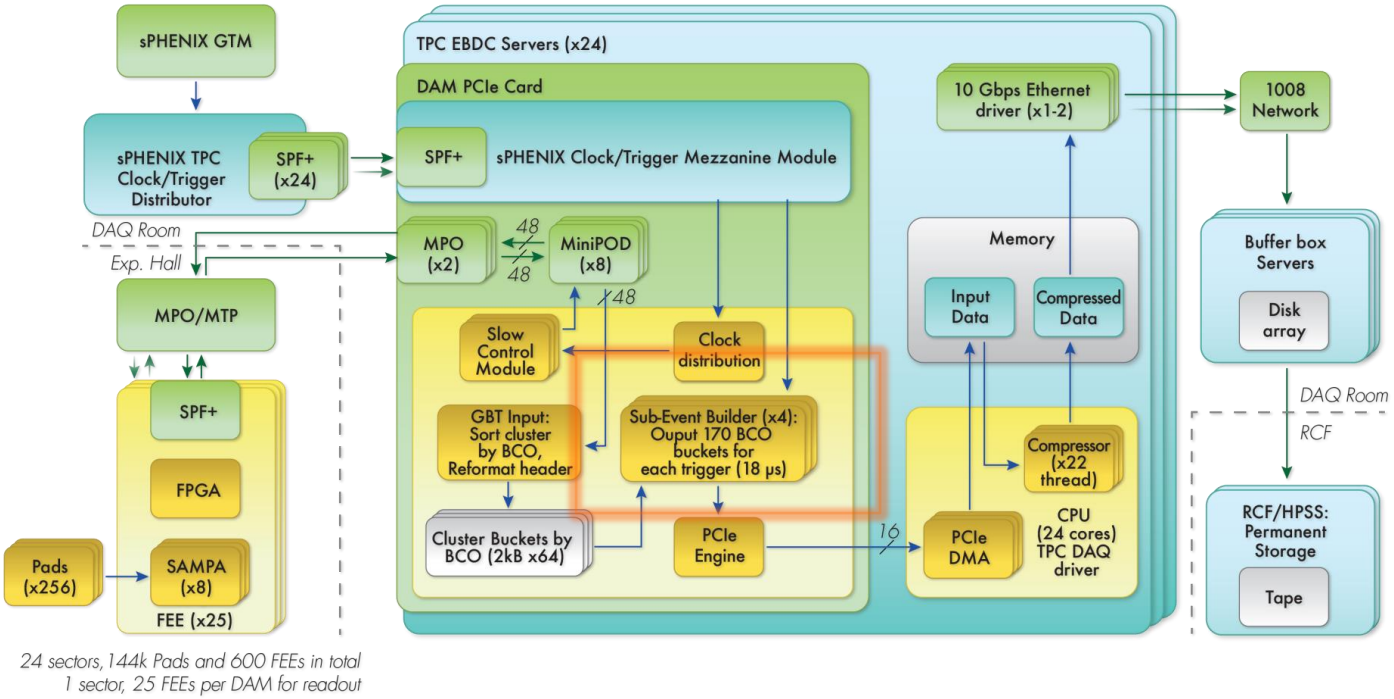
BCO buckets data format

- 25 DAM channels each handle one FEE
- Each channel buffer clusters sorted in 64 BCO buckets, based on hit time of the first sample of cluster
- Require 25 (FEE) x64 (BCO) memory block or FIFO, each store one BCO buckets
- Byte aligned
- Overhead = 9x8bit / 5x10 bit = 144%

One BCO buckets per FEE per BCO
Max 255 clusters



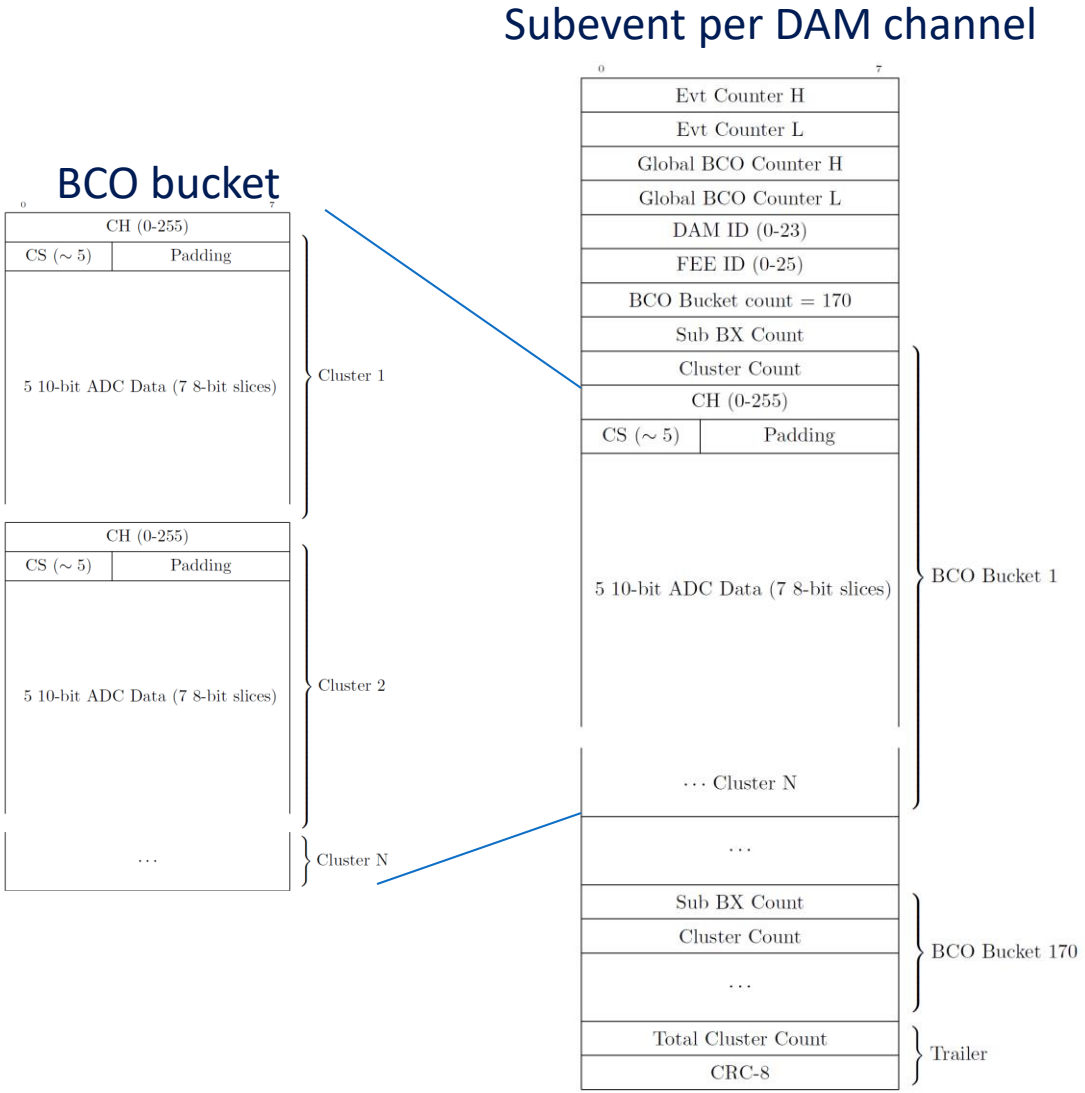
Throttling VS triggering



- 15kHz trigger + 170 BCO readout length (readout 18us data per trigger) → only need ~25% data from the input continues stream
- Two options
  - Throttling: only record hits within 170BCO of the trigger and form a continuous data stream; no duplicated hits. **Data reduction to 25.5%**
  - Trigger: for each trigger, readout a chunk of hits timed to the next 170BCO. Form sub-event and easy for analysis; but could duplicate hits in output data if two trigger comes within 170 BCO. **Data reduction to 28.5%**
- Since the trigger mode only increase data volume by 10% (relatively), I would prefer trigger mode instead of throttled mode for easy analysis and monitoring.
- Output average continuous rate = 7 Gbps **(30 pad rows, reduced from 25 Gbps)**

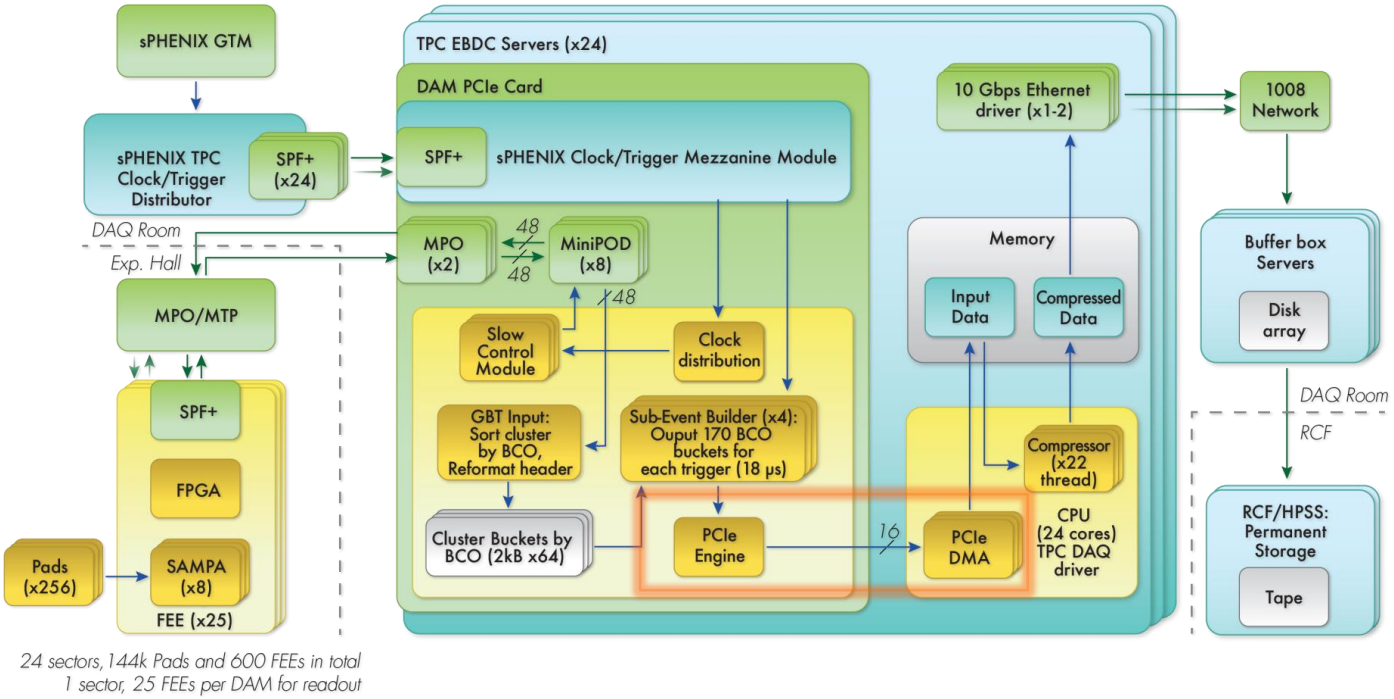
Event builder data format

- Each DAM channel assemble data for all 170 buckets. One DAM channel per FEE
- Add header and trailer
- Send to memory via PCIe engine simultaneously for 25 DAM channels?



FPGA -> CPU

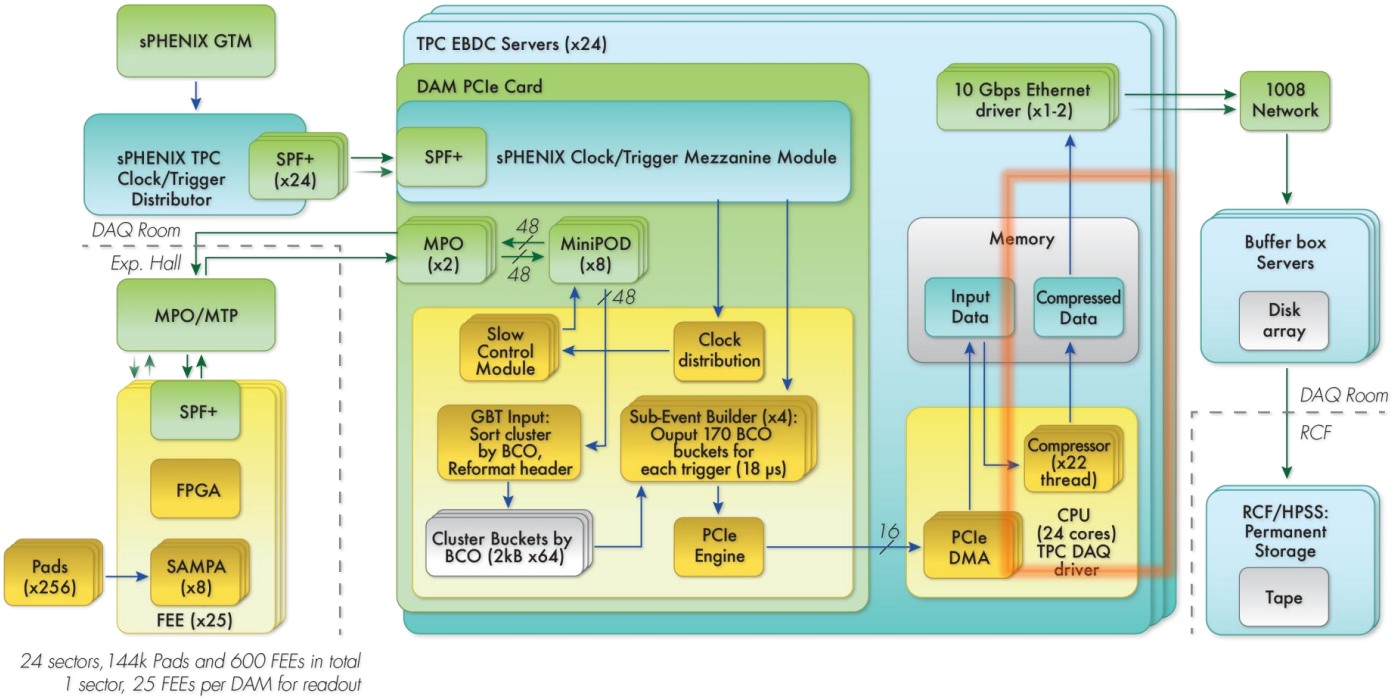
Rate estimation spread sheets:
https://docs.google.com/spreadsheets/d/1Q_uYf00_8pushSiYns29T_-ThIOqQaqpKbVS_LDqIAg/edit?usp=sharing



- Merge data from 25 DAM/FEE channels and their multiple event builders to PCIe engine for output (need detail design)
- FIFO and DMA event building output to Server Memory
  - Media: PCIe Gen3 x16
- Demonstrated rate limit for FELIX (PCIe x16) ~ 100 Gbps
- Average continuous rate = 7 Gbps **(30 pad rows)**

Data compression

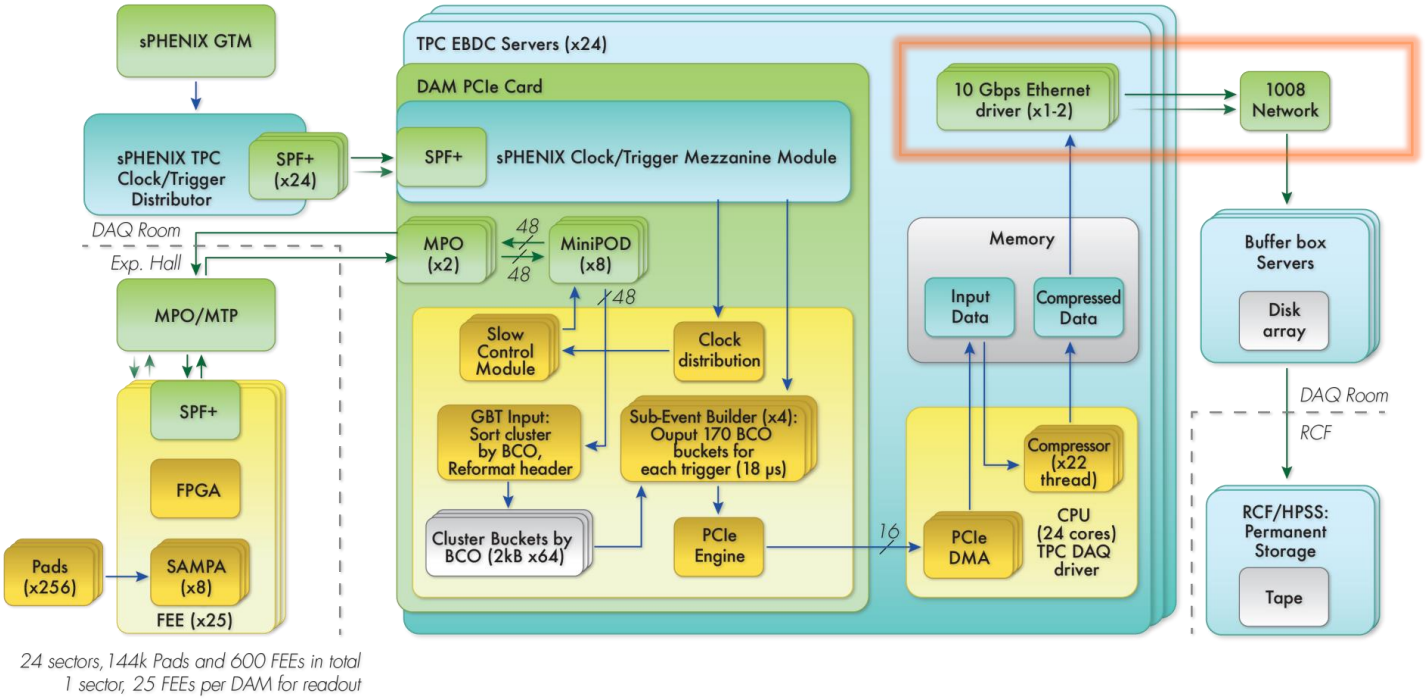
Rate estimation spread sheets:
https://docs.google.com/spreadsheets/d/1Q_uYf00_8pushSiYns29T_-ThIOqQaqpKbVS_LDqIAg/edit?usp=sharing



- Multithread compression
  - Algorithm: LZO on multi-event chunks
  - Demonstrated PHENIX data compression ratio = 60%, need to emulate for TPC data
- **Estimated rate limit = 60 MBps / core x (10-20) core >= 4.8 Gbps**
- Average continuous rate = 4.2 Gbps **(30 pad rows)**

Output stage

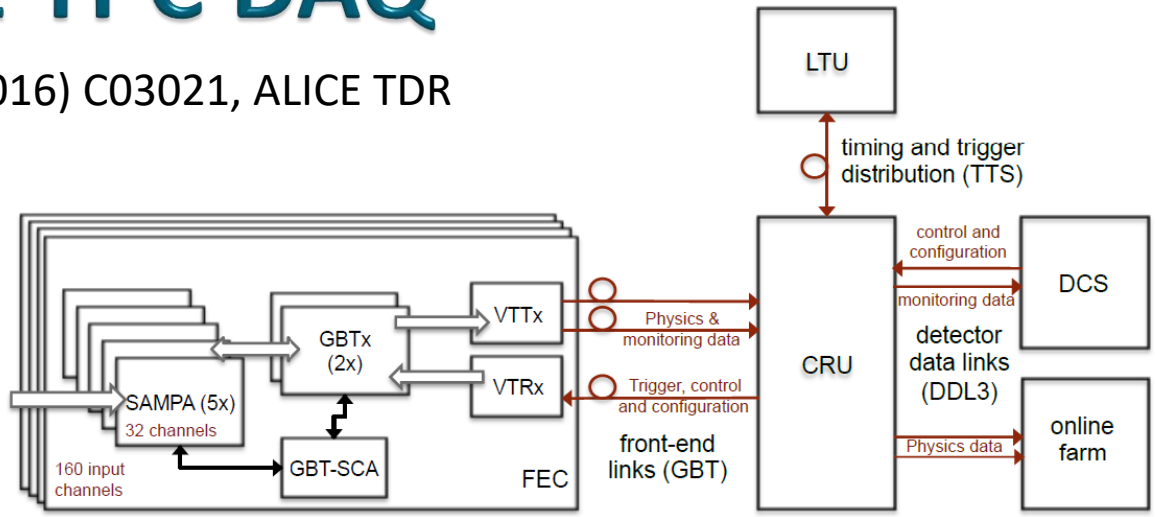
Rate estimation spread sheets:
https://docs.google.com/spreadsheets/d/1Q_uYf00_8pushSiYns29T_-ThIOqQaqpKbVS_LDqIAg/edit?usp=sharing



- Output to event builder
  - Media: 1x 10 Gbps Ethernet ports per EBDC server
- Rate limit buffer box = 120 Gbps total? (3x HPSS rate)
- Average continuous rate for whole system
  = 4.2 Gbps/EBDC * 24 EBDC = 100 Gbps **(30 pad rows)**
  - **Very close to 120 Gbps limit!**

ALICE TPC DAQ

JINST 11 (2016) C03021, ALICE TDR

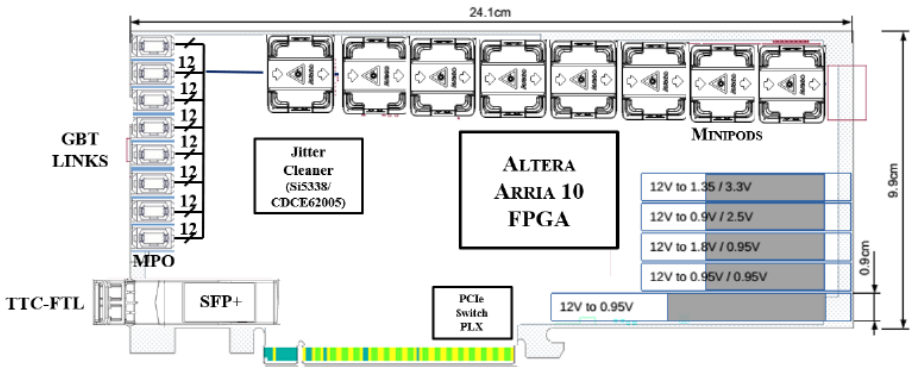


(a) PCIe40.

Figure 6.9: Schematic of the TPC readout system with the CRU as central part interfacing the front-end electronics to the trigger system, the DCS and the online farm.

ALICE CRU based on LHCb PCIe40 card

- Prototyped by CPPM, Marseille, France
- Arria 10 family FPGA, (15K$/chip?)
- 24 GBT input fibers [JINST 11 2016]
- PCIe Gen3 x16 interface
- TTC-FTL accepting timing/trigger
- Cost 15-20 k$ (need to be confirmed)

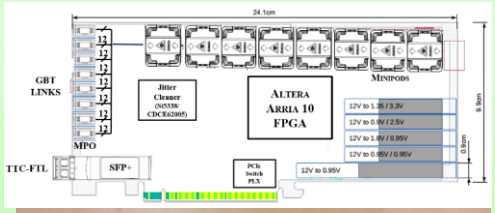


(b) PCIe40 Schematic.

Our options: 10-50x (PCIe card + server)

Data Aggregation Module (DAM):
PCIex8 or x16 card with multiple (8-48x) GBT fiber IO

Option 1: LHCb/ALICE CRU

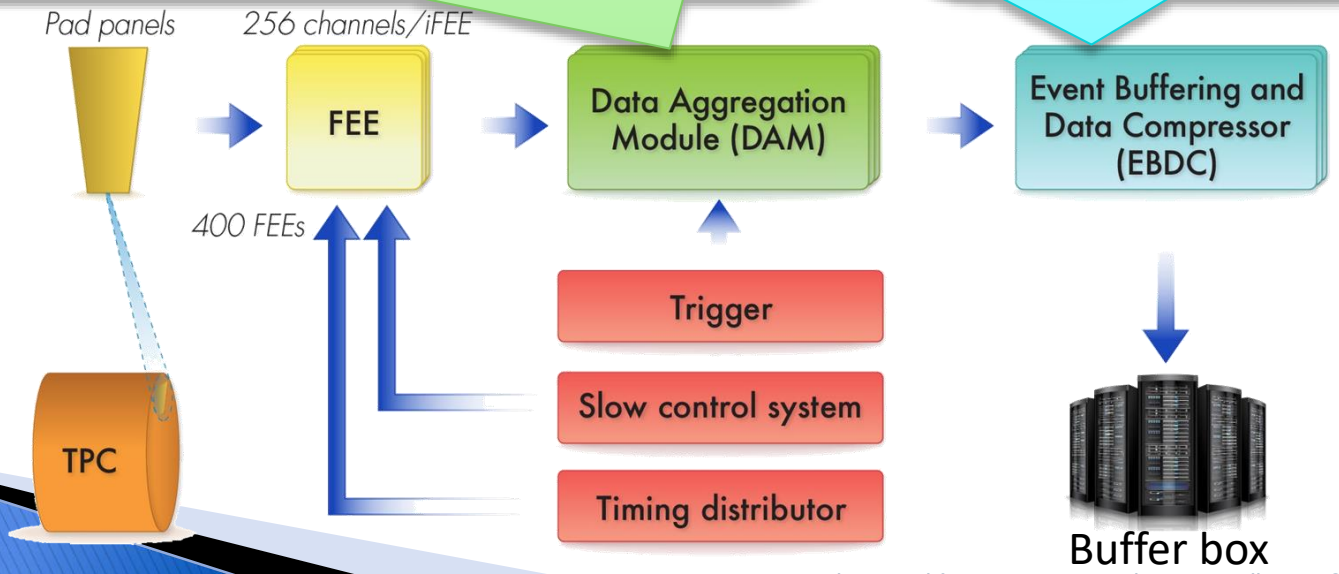


Option 2: ATLAS FELIX
(see next talk)

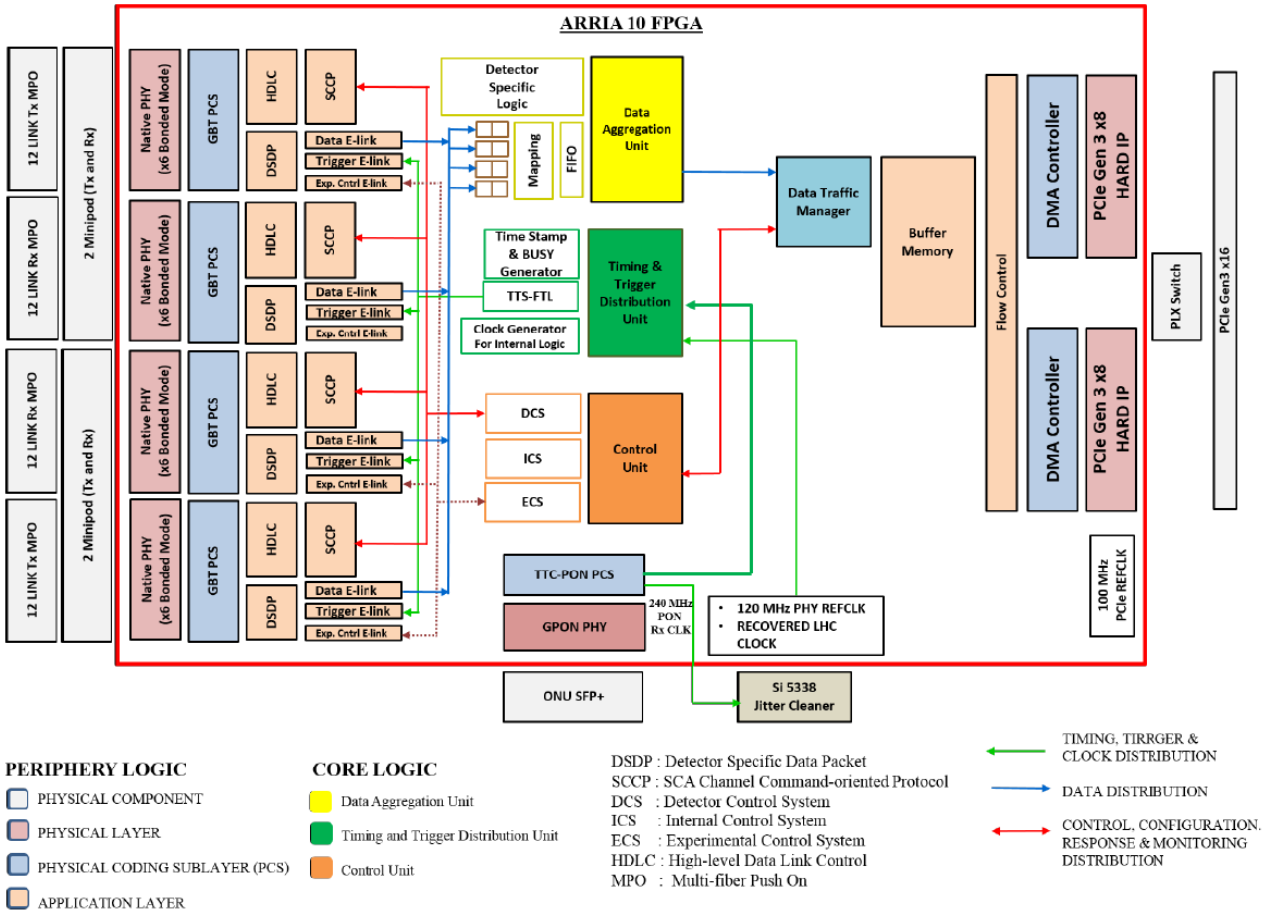
Option 3: build our own based on ALICE/ATLAS exp.

Event Buffering and Data Compressor (EBDC): Rack server that can host at 1x PCIex16 cards + 2x 10 Gbps Ethernet port

Example: Dell PowerEdge R830
2x12 cores, 2x10 GBps, ~ 10k$



CRU diagram


ARRIA 10 FPGA
12 LINK Tx MPO
12 LINK Rx MPO
2 Minipod (Tx and Rx)
12 LINK Rx MPO
12 LINK Tx MPO
2 Minipod (Tx and Rx)
Native PHY (x6 Bonded Mode)
GBT PCS
HDLC
DSDP
SCCP
Data E-link
Trigger E-link
Exp. Cntrl E-link
Detector Specific Logic
Mapping
FIFO
Data Aggregation Unit
Time Stamp & BUSY Generator
TTS-FTL
Clock Generator For Internal Logic
Timing & Trigger Distribution Unit
DCS
ICS
ECS
Control Unit
TTC-PON PCS
GPON PHY
240 MHz PON Rx CLK
ONU SFP+
Si 5338 Jitter Cleaner
120 MHz PHY REFCLK
RECOVERED LHC CLOCK
Data Traffic Manager
Buffer Memory
Flow Control
DMA Controller
PCIe Gen 3 x8 HARD IP
PLX Switch
PCIe Gen3 x16
100 MHz PCIe REFCLK
PERIPHERY LOGIC
PHYSICAL COMPONENT
PHYSICAL LAYER
PHYSICAL CODING SUBLAYER (PCS)
APPLICATION LAYER
CORE LOGIC
Data Aggregation Unit
Timing and Trigger Distribution Unit
Control Unit
DSDP : Detector Specific Data Packet
SCCP : SCA Channel Command-oriented Protocol
DCS : Detector Control System
ICS : Internal Control System
ECS : Experimental Control System
HDLC : High-level Data Link Control
MPO : Multi-fiber Push On
TIMING, TIRRGER & CLOCK DISTRIBUTION
DATA DISTRIBUTION
CONTROL, CONFIGURATION, RESPONSE & MONITORING DISTRIBUTION

SAMPA/STAR iFEE

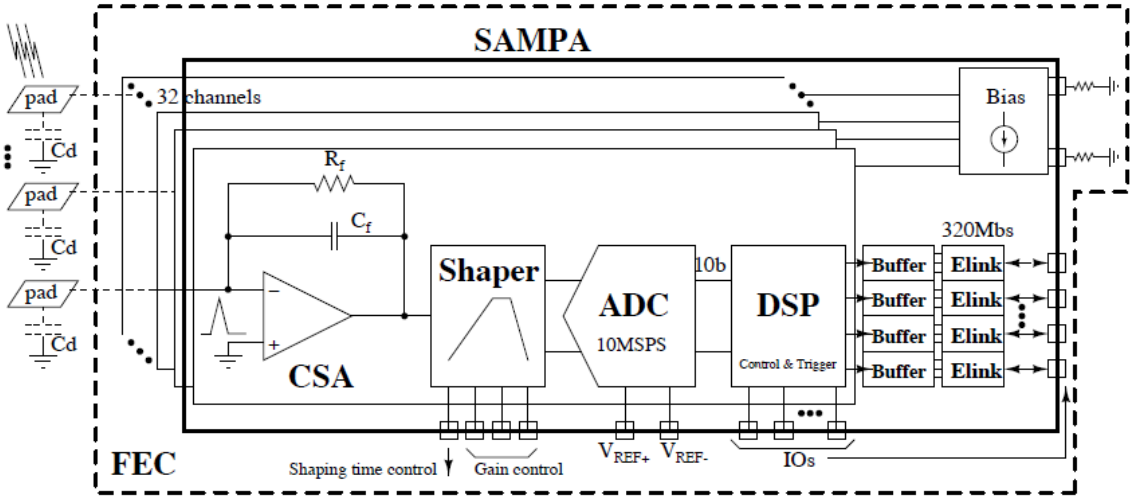
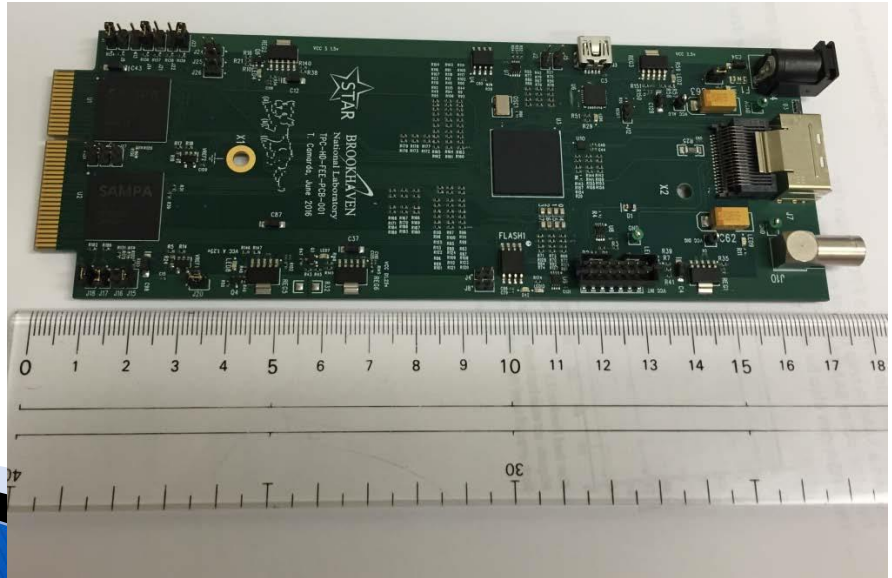


Figure 6.4: Schematic of the SAMPA ASIC for the GEM TPC readout, showing the main building blocks.



Sept 2016 cost estimate

Cost estimate (for production)

- Direct M&S cost for 100K channels is 1.1M FY16$
- Cost for development is not included.
  - We assumed ~20-30% of this as M&S cost for development

Item	# of items	$ per item	$ all
SAMPA Chips	3200	$44	$140K
FEE cards	400	$700	$280K
DAM	50	$6000	$300K
Cables/fibers			$100K
Power Supply	8	$12000	$100K
EBDC	50	$3000	$150K
Total			**$1.1M**

c.f. STAR iFEE is $150/card (64 ch., copper cable readout)